When Genes Speak: A Semantic-Guided Framework for Spatially Resolved Transcriptomics Data Clustering

Jiangkai Long[1], Yanran Zhu[1], Chang Tang[2*] Kun Sun[1], Yuanyuan Liu[1], Xuesong Yan[1]

[1]School of Computer Science, China University of Geosciences, Wuhan 430074, China
[2]School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
longjk@cug.edu.cn, tangchang@hust.edu.cn

Abstract

Spatial transcriptomics enables gene expression profiling with spatial context, offering unprecedented insights into the tissue microenvironment. However, most computational models treat genes as isolated numerical features, ignoring the rich biological semantics encoded in their symbols. This prevents a truly deep understanding of critical biological characteristics. To overcome this limitation, we present SemST, a semantic-guided deep learning framework for spatial transcriptomics data clustering. SemST leverages Large Language Models (LLMs) to enable genes to "speak" through their symbolic meanings, transforming gene sets within each tissue spot into biologically informed embeddings. These embeddings are then fused with the spatial neighborhood relationships captured by Graph Neural Networks (GNNs), achieving a coherent integration of biological function and spatial structure. We further introduce the Fine-grained Semantic Modulation (FSM) module to optimally exploit these biological priors. The FSM module learns spot-specific affine transformations that empower the semantic embeddings to perform an element-wise calibration of the spatial features, thus dynamically injecting high-order biological knowledge into the spatial context. Extensive experiments on public spatial transcriptomics datasets show that SemST achieves state-of-the-art clustering performance. Crucially, the FSM module exhibits plug-and-play versatility, consistently improving the performance when integrated into other baseline methods.

Code — https://github.com/longjiangk/SemST

Introduction

Understanding the intricate spatial architecture of tissues is a cornerstone of modern biology and medicine, holding the key to deciphering mechanisms of organogenesis, disease progression, and therapeutic response (Ståhl et al. 2016; Longo et al. 2021). The advent of spatial transcriptomics has revolutionized this pursuit by enabling the measurement of gene expression profiles while preserving their native spatial coordinates (Rodriques et al. 2019; Marx 2021). A fundamental computational task in spatial transcriptomics is the identification of spatial domains—regions with similar gene expression patterns that often reflect an underlying anatomical or functional organization (Biancalani et al. 2021). Robust spatial domain identification lays a solid foundation for downstream biological analyses, such as differential gene expression analysis, functional enrichment analyses, and cell–cell interaction inference (Dong and Zhang 2022). This task is commonly formulated as a clustering problem, aiming to partition the tissue into biologically meaningful units (Du et al. 2023).

Recent years have seen a surge in deep learning models, particularly those based on Graph Neural Networks (GNNs) (Li et al. 2022; Liu et al. 2024b; Zahedi et al. 2024). These methods have achieved considerable success by modeling the tissue as a graph, where spots (or cells) are nodes and spatial proximity defines the edges. By aggregating information from neighboring spots, GNNs can effectively learn representations that capture the spatial context of gene expression. Another advantage is that the graph topology naturally reflects possible cell–cell interactions, offering an implicit yet biologically meaningful view of intercellular communication (Zhou 2022; Lampert et al. 2024).

However, a critical and pervasive limitation persists: these models overwhelmingly treat genes as isolated, high-dimensional numerical features. This approach completely ignores the vast repository of biological knowledge encapsulated in the gene symbols themselves—knowledge about their functions, pathways, and interactions. For example, "Postn" has been implicated in extracellular matrix remodeling and fibroblast activation, whereas the co-expression of "Ttn" and "Actc1" is indicative of cardiac muscle identity. This semantic gap prevents the models from reasoning about the tissue in a biologically informed manner, thereby limiting their capacity for deeper biological discovery.

The recent success of general-purpose Large Language Models (LLMs) in bioinformatics (Liu et al. 2024a) has opened new possibilities for leveraging textual gene knowledge in computational biology. For instance, GenePT (Chen and Zou 2025) leverages ChatGPT (Brown et al. 2020) to generate embeddings for individual genes based on their textual descriptions or symbols, facilitating diverse applications in gene and single-cell analysis. In the context of integrating transcriptomics with histopathology, models such as SGN (Yang et al. 2024) and OmiCLIP (Chen et al. 2025) adopt a similar approach, generating gene semantic em-

*Corresponding author

beddings alongside image embeddings, which enables joint modeling of gene symbols and tissue morphology. These works demonstrate the feasibility of translating gene symbols into biologically meaningful concepts. However, despite its promise, the direct application of this semantic-aware approach to spatial transcriptomics data clustering has remained challenging and largely unexplored. The primary challenge lies in the fundamental disparity between symbolic biological knowledge and quantitative spatial information, two modalities with inherently different characteristics that are difficult to reconcile.

To bridge this gap, we introduce SemST, a novel semantic-guided framework for spatial transcriptomics data clustering. SemST explores the integration of gene semantics directly into the spatial modeling process. We leverage the power of LLMs, pre-trained on vast biomedical text corpora, to transform discrete gene symbols into dense, continuous embeddings that map spots within a meaningful biological semantic space. The central challenge then becomes how to effectively integrate these powerful semantic priors with the spatial relationship features. A naive fusion risks diluting the rich information from each modality, rather than creating a synergistic whole.

To this end, we introduce the Fine-grained Semantic Modulation (FSM) module. Instead of passively combining information, the FSM module enables an active, dynamic "dialogue" between biological functions and spatial contexts. Specifically, the FSM learns a spot-specific affine transformation to perform an element-wise calibration of the GNN-derived spatial features. This mechanism allows high-order biological knowledge to dynamically guide and refine the representation of local spatial context, leading to more robust and biologically plausible representations.

The contributions of this work can be summarized as the following three-fold:

- We propose SemST, a novel deep learning framework that effectively fuses LLM-derived gene semantics with GNN-captured spatial context for superior spatial domain identification, moving beyond the conventional treatment of genes as isolated numerical features.
- We design the Fine-grained Semantic Modulation (FSM) module, which enables semantic priors to dynamically calibrate spatial features in an element-wise manner, resulting in the infusion of high-order biological knowledge into spatial representations.
- Extensive experiments on multiple benchmark spatial transcriptomics datasets demonstrate that SemST significantly outperforms existing state-of-the-art methods. We also show that our proposed module is a versatile, plug-and-play component that consistently boosts the performance of other baseline models.

Related Work

Large Language Models for Biological Semantics

Recent studies have demonstrated the potential of LLMs in interpreting biological data by treating gene symbols as natural language. For instance, models like GPT-4 can perform zero-shot cell type annotation with remarkable accuracy by simply inputting lists of marker genes (Hou and Ji 2024; Zeng and Du 2023). In addition, LLMs are capable of generating coherent biological summaries and GO (Ashburner et al. 2000) terms, often rivaling the performance of human experts (Huang 2024; Hu et al. 2025). From the perspective of semantic embeddings, GenePT (Chen and Zou 2025) demonstrates the capability of GPT models to generate informative embeddings from either gene description texts or gene symbols alone. When integrating transcriptomic data with histopathological images, SGN (Yang et al. 2024) performs dot-product operations between LLM-derived gene embeddings and visual features, enabling zero-shot gene expression prediction. Similarly, OmiCLIP (Chen et al. 2025) employs a contrastive learning framework to align gene semantics with image patch features within a unified feature space, thereby establishing a visual-omics foundation model. These methods all successfully leverage the language processing power of LLMs to unlock biological insights, demonstrating the potential of incorporating linguistic information into spatial transcriptomic data analysis.

Spatial Domain Identification

The primary goal of spatial domain identification in spatial transcriptomics is to cluster spots with similar characteristics, a task crucial for understanding tissue architecture. Early approaches such as Louvain (Blondel et al. 2008) and Seurat (Satija et al. 2015) typically overlooked spatial context. To address this, methods specifically designed to incorporate spatial information have emerged. BayesSpace (Zhao et al. 2021) employs a Bayesian model with a spatial prior to encourage neighboring spots to belong to the same cluster. SpaGCN (Hu et al. 2021) was a pioneering deep learning approach, constructing a graph from the tissue and using a Graph Convolutional Network (GCN) to integrate gene expression with spatial coordinates. Building on this, a family of GNN-based models has become a dominant paradigm. STAGATE (Dong and Zhang 2022) uses a graph attention autoencoder to learn low-dimensional embeddings that capture both expression similarity and spatial proximity. GraphST (Long et al. 2023) further enhances this by incorporating contrastive learning to better distinguish between different spatial domains. Spatial-MGCN (Wang et al. 2023) employs a multi-view GCN to model the complex relationships between gene expression and spatial information. However, a limitation persists across all these methods: they treat gene expression as a numerical matrix, discarding the rich biological knowledge encoded in gene symbols.

Method

In this section, we present SemST, a novel semantic-guided deep learning framework for spatial transcriptomics data clustering. The overall architecture of SemST is illustrated in Figure 1.

Problem Formulation

We consider a spatial transcriptomics dataset denoted as $\mathcal{D} = \{\mathbf{X}, \mathbf{S}, \mathcal{G}\}$. The normalized gene expression matrix is

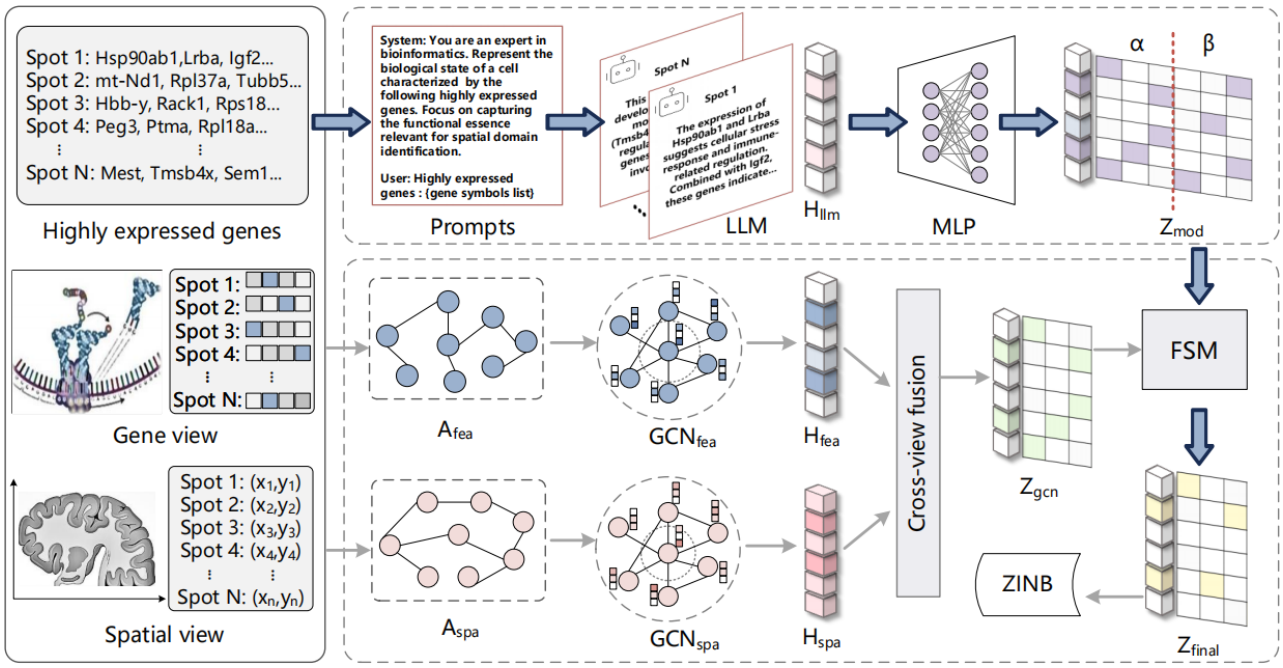


Figure 1: The framework of the proposed SemST. We integrate three distinct information as inputs. Two graphs are constructed based on expression and spatial data, and multi-view features are propagated and fused using GCN. Gene symbols are formatted into prompts and fed into an LLM to extract biologically informed embeddings, which are then used by the proposed FSM module to guide spatial feature refinement. Model optimization is achieved through data reconstruction based on ZINB.

represented by $\mathbf{X} \in \mathbb{R}^{N \times M}$, where N is the number of spots and M is the number of genes. The spatial coordinates of the spots are given by $\mathbf{S} \in \mathbb{R}^{N \times 2}$. The set of gene symbols is denoted by $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$. The primary objective of our framework is to learn a low-dimensional feature representation for each spot, such that these representations can be effectively utilized for clustering to reveal underlying spatial domains and functional regions within the tissue.

Multi-View GNN for Spatial Representation

To comprehensively capture the relationships among spots, we construct two distinct graphs: one based on spatial proximity and the other on gene expression similarity. The former excels at modeling microenvironmental interactions, while the latter is well suited for identifying spots that belong to the same layer but are spatially distant.

Spatial Proximity Graph. We construct a spatial graph $\mathbf{A}_{spa} \in \{0,1\}^{N \times N}$ based on the physical coordinates $\mathbf{S}$. An edge is established between two spots i and j if their Euclidean distance is within a predefined radius threshold r. The adjacency matrix is formally defined as:

$$(\mathbf{A}_{spa})_{ij} = \begin{cases} 1, & \text{if } \|\mathbf{S}_i - \mathbf{S}_j\|_2 \leq r, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Gene Expression Similarity Graph. We construct a feature graph $\mathbf{A}_{fea} \in \{0,1\}^{N \times N}$ using the K-Nearest Neighbors (KNN) (Cover and Hart 1967) algorithm on the gene expression matrix $\mathbf{X}$. An edge connects spot i to spot j if spot j is among the top k_n most similar spots to spot i in the high-dimensional expression space. This is defined as:

$$(\mathbf{A}_{fea})_{ij} = \begin{cases} 1, & \text{if } j \in \mathcal{N}(i), \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{N}(i)$ represents the set of k_n nearest neighbors of spot i based on the cosine similarity of their expression profiles $\mathbf{X}_i$ and $\mathbf{X}_j$.

Multi-View GCN Propagation. With the two graphs constructed, we employ a multi-view GCN to learn latent representations from both spatial and feature perspectives. Each branch processes the gene expression matrix $\mathbf{X}$ as node features but uses a different graph structure.

The propagation rule for a GCN layer is given by:

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), \tag{3}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with self-loops, $\tilde{\mathbf{D}}$ is the corresponding degree matrix, $\mathbf{H}^{(l)}$ is the feature matrix at layer l ($\mathbf{H}^{(0)} = \mathbf{X}$), $\mathbf{W}^{(l)}$ is a trainable weight matrix at layer l, and σ is a non-linear activation function.

The two branches generate embeddings as follows:

$$\mathbf{H}_{spa} = \text{GCN}_{spa}(\mathbf{X}, \mathbf{A}_{spa}), \tag{4}$$

$$\mathbf{H}_{fea} = \text{GCN}_{fea}(\mathbf{X}, \mathbf{A}_{fea}), \tag{5}$$

where GCN_{spa} and GCN_{fea} are multi-layer GCNs with separate parameters. The fused spatial representation, $\mathbf{Z}_{gcn} \in \mathbb{R}^{N \times d}$, is obtained by concatenating the outputs of the two branches:

$$\mathbf{Z}_{gcn} = \text{Concat}(\mathbf{H}_{spa}, \mathbf{H}_{fea}). \tag{6}$$

This multi-view approach allows the model to learn a robust representation that jointly captures local spatial proximity and distance-independent expression profile similarity.

Semantic-Guided Feature Modulation

Fundamental to SemST is its ability to leverage the rich biological knowledge inherent in gene symbols. We achieve this by generating semantic embeddings using a general LLM and then using these embeddings to modulate the GNN-derived representation $\mathbf{Z}_{gcn}$ through our FSM module.

Gene Semantics Extraction. For each spot i, we begin by identifying its key biological characteristics through the selection of the top k_g most highly expressed genes, denoted as $\mathcal{G}_i^k = \{g_{i,1}, g_{i,2}, \ldots, g_{i,k_g}\}$. These gene symbols are formatted into a natural language string and incorporated into a descriptive prompt, which is subsequently fed into an LLM. The parameters of LLM remain frozen throughout training, allowing it to serve as a fixed, comprehensive biological knowledge base. From the LLM, we extract the final-layer hidden state, denoted as $\mathbf{H}_{llm} \in \mathbb{R}^{N \times d'}$, which encodes rich, high-order biological knowledge.

Fine-grained Semantic Modulation (FSM). Integrating semantic and spatial information is crucial for comprehensive biological understanding. While conventional fusion strategies (e.g., concatenation, addition, cross-attention) are commonly employed, they often exhibit limited efficacy in fully leveraging the biological prior knowledge from the LLM. A more fine-grained and expressive modulation is required to facilitate semantic-spatial integration (Perez et al. 2018). The proposed FSM module specifically aims to adapt the powerful LLM-derived biological prior for our task. Firstly, we project the $\mathbf{H}_{llm}$ through a trainable Multi-Layer Perceptron (MLP). This MLP transforms the high-dimensional $\mathbf{H}_{llm}$ into a task-specific modulation space, yielding the semantic modulation matrix $\mathbf{Z}_{mod} \in \mathbb{R}^{N \times 2d}$ with an output dimension of $2 \times d$:

$$\mathbf{Z}_{mod} = f_{mlp}(\mathbf{H}_{llm}). \tag{7}$$

Subsequently, splitting $\mathbf{Z}_{mod}$ along its feature dimension into two equal-sized tensors: a scaling factor $\boldsymbol{\alpha} \in \mathbb{R}^{N \times d}$ and a bias factor $\boldsymbol{\beta} \in \mathbb{R}^{N \times d}$:

$$[\boldsymbol{\alpha} \mid \boldsymbol{\beta}] = \mathbf{Z}_{mod}. \tag{8}$$

The final, semantically-calibrated representation $\mathbf{Z}_{final}$ is computed as:

$$\mathbf{Z}_{final} = ((1 + \boldsymbol{\alpha}) \odot \mathbf{Z}_{gcn} + \boldsymbol{\beta})\mathbf{W}, \tag{9}$$

where $\odot$ denotes the element-wise (Hadamard) product, and $\mathbf{W}$ is a trainable weight matrix. This affine transformation enables the semantic representation to dynamically modulate each spot's spatial features through fine-grained scaling and shifting, effectively leveraging the biological knowledge captured by the LLM to actively steer the representation—letting the genes "speak" in the learning process.

Notably, we intentionally generate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from the same MLP to ensure they are embedded in the same bio-semantic space. During backpropagation, the distinct gradients for scaling and shifting lead them to gradually differentiate and specialize, effectively decoupling identical semantics into distinct functions. Using two separate networks to generate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ risks uncoordinated modulation.

Model Optimization

To train the entire framework without ground-truth labels, we adopt a self-supervised learning strategy based on data reconstruction.

ZINB Loss. Given the sparse and over-dispersed nature of spatial transcriptomics data, we use a Zero-Inflated Negative Binomial (ZINB) reconstruction loss (Yu et al. 2022). The decoder network transforms the final embedding $\mathbf{Z}_{final}$ into the three parameter matrices of the ZINB distribution for each gene in each spot: the mean $\boldsymbol{\mu} = f_\mu(\mathbf{Z}_{final})$, the dispersion $\boldsymbol{\theta} = f_\theta(\mathbf{Z}_{final})$, and the dropout probability $\boldsymbol{\pi} = f_\pi(\mathbf{Z}_{final})$. The ZINB loss is the negative log-likelihood of the observed expression counts $\mathbf{X}$ given these parameters:

$$\mathcal{L}_{zinb} = -\sum_{i=1}^{N}\sum_{j=1}^{M} \log P(\mathbf{X}_{ij} \mid \boldsymbol{\mu}_{ij}, \boldsymbol{\theta}_{ij}, \boldsymbol{\pi}_{ij}). \tag{10}$$

Furthermore, we adopt two additional loss functions, each of which has been demonstrated to be highly effective in prior studies (Zhu et al. 2024; He et al. 2024).

Correlation Reduction Loss. To strengthen the GNN backbone, we introduce correlation reduction loss. Specifically, given two embeddings $\mathbf{H}_{spa}$ and $\mathbf{H}_{fea}$, we compute the cosine similarity between the i-th feature in spatial view and the j-th feature in the feature view:

$$\mathbf{C}_{ij} = \frac{(\mathbf{H}_i^{spa})(\mathbf{H}_j^{fea})^\top}{\|\mathbf{H}_i^{spa}\| \, \|\mathbf{H}_j^{fea}\|}. \tag{11}$$

Then, the correlation reduction loss encourages correlation matrix $\mathbf{C} \in \mathbb{R}^{p \times p}$ (where $p = d/2$) to be close to the identity matrix $\mathbf{I}$, thereby promoting consistence of corresponding features across views while reducing correlations between distinct features. The loss is formulated as:

$$\mathcal{L}_{cr} = \frac{1}{p^2}\sum_{i,j}(\mathbf{C}_{ij} - \mathbf{I}_{ij})^2. \tag{12}$$

Spatial Regularization Loss. We introduce a spatial regularization loss to prevent the over-injection of semantics that may harm spatial coherence. This loss ensures that spatially neighboring spots are close, while non-neighboring spots are distant in the latent space $\mathbf{Z}_{final}$, formulated as:

$$\mathcal{L}_s = -\sum_{i=1}^{N}\left(\sum_{j \in \mathcal{R}_i} \log \sigma(\psi_{ij}) + \sum_{k \notin \mathcal{R}_i} \log\left(1 - \sigma(\psi_{ik})\right)\right), \tag{13}$$

where ψ_{ij} denotes the cosine similarity between the representations of spots i and j, $\sigma(\cdot)$ denotes the sigmoid function and $\mathcal{R}_i$ represents the set of spatial neighbors of spot i.

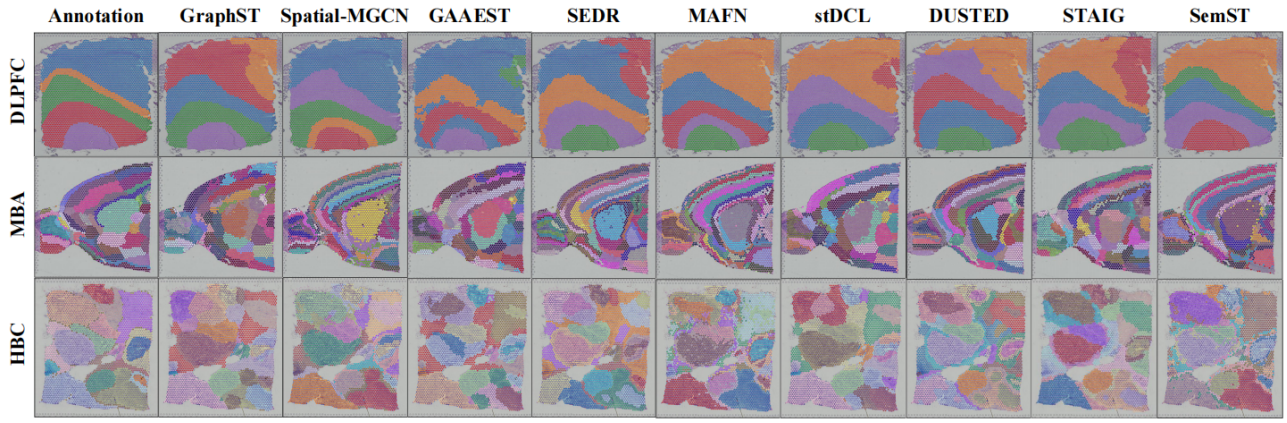


Figure 2: Visualization of manual annotations and clustering results produced by SemST and eight other methods on the DLPFC slice #151672, MBA, and HBC datasets. Color indicates spatial domains.

Overall Loss Function. The total loss function for SemST is formulated as:

$$\mathcal{L} = \mathcal{L}_{zinb} + \gamma \mathcal{L}_{cr} + \lambda \mathcal{L}_{s}, \tag{14}$$

where γ and λ are hyper-parameters that balance the contributions of the different loss terms.

Experiments

Experimental Setup

Datasets. We evaluate our method on nine public spatial transcriptomics datasets. The human dorsolateral prefrontal cortex dataset (DLPFC), the human breast cancer dataset (HBC), and the mouse brain anterior dataset (MBA) are all generated using the 10x Genomics Visium platform. For the DLPFC dataset, we use five tissue slices—151508, 151509, 151510, 151671, and 151672—obtained from two adult human samples, with each section manually annotated into 5 or 7 cortical layers. The HBC dataset is annotated into 20 spatial domains, and the MBA dataset contains 52 annotated spatial domains. We also include the mouse embryo dataset (ME) from the Stereo-seq platform, which is derived from an E9.5 sample and includes 12 annotated spatial domains, and the mouse visual cortex dataset (MVC) from the STARmap platform, which contains 7 annotated spatial domains.

Compared Methods. We compare our method with nine representative methods, including STAGATE (Dong and Zhang 2022), GraphST (Long et al. 2023), Spatial-MGCN (Wang et al. 2023), GAAEST (Wang et al. 2024), SEDR (Xu et al. 2024), MAFN (Zhu et al. 2024), stDCL (Yu et al. 2025), DUSTED (Zhu et al. 2025), and STAIG (Yang et al. 2025). All baseline results are obtained by running official implementations with recommended parameters to achieve the best performance.

Metrics. We adopt four standard clustering metrics to evaluate performance: Adjusted Rand index (ARI), Normalized Mutual Information (NMI), clustering accuracy (ACC), and F1-score (F1). These metrics comprehensively assess the agreement between the predicted clusters and ground-truth annotations. We multiplied all the results by 100.

Implementation Details. Our model is implemented in PyTorch and trained using the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 5×10^{-4}. All experiments are conducted on a machine equipped with an NVIDIA RTX 3090 GPU (24GB). More implementation details can be found in the supplementary materials.

Quantitative and Qualitative Results

Table 1 presents the quantitative clustering results on eight spatial transcriptomics datasets. SemST almost consistently outperforms all compared methods across most metrics and datasets. For example, on the 151508 slice, SemST surpasses the closest competitor by over 10 ARI points. On the MBA dataset, characterized by highly complex spatial structures that make distinguishing small domains particularly challenging, SemST achieves an ARI improvement of more than 5 points over the best compared method.

Complementing these quantitative findings, Figure 2 provides a qualitative visualization of spatial clustering results on three representative datasets. The visualizations clearly show that SemST produces more anatomically coherent and biologically meaningful clusters, aligning closely with annotated ground truth spatial domains. These results demonstrate that SemST effectively captures both functional and structural tissue patterns, achieving superior accuracy and biological consistency.

Ablation Study

To systematically evaluate the contribution of the key architectural components of SemST and validate the versatility of the FSM module, we conducted comprehensive ablation experiments. Table 2 presents the performance of SemST variants on four representative datasets. Table 3 shows the clustering performance improvement achieved by integrating the FSM module into all compared methods.

Absence of Semantic Embeddings. The "w/o LLM" variant represents SemST trained without LLM-derived semantic embeddings. As shown, removing the semantic embeddings leads to a performance drop across all metrics and

Datasets	151508				151509				151510				151671			
Methods	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1
STAGATE[Nat. Com.'22]	53.84	64.10	69.39	64.13	49.54	65.81	67.42	64.53	46.07	59.17	61.55	50.88	59.54	69.37	75.10	77.56
GraphST[Nat. Com.'23]	48.61	64.28	61.93	56.78	51.93	67.41	67.48	63.40	51.29	64.53	68.38	53.35	61.02	71.77	74.84	79.62
Spatial-MGCN(BIB'23)	46.22	60.23	59.46	56.45	54.22	68.16	63.83	63.27	51.61	66.50	59.35	52.38	60.19	71.87	73.39	75.18
GAAEST[Com. Biol.'24]	31.24	52.82	57.18	44.79	43.84	53.16	67.21	43.39	39.79	53.21	59.61	45.30	64.80	65.89	74.47	71.41
SEDR[Genome Med.'24]	47.47	62.07	61.79	60.53	49.69	63.30	66.54	58.36	50.52	64.76	60.20	60.70	59.97	68.25	74.44	76.96
MAFN[TKDE'24]	54.54	66.31	66.26	62.61	70.48	69.41	75.36	66.07	69.91	70.67	74.82	64.07	72.41	68.05	79.53	79.29
stDCL[Adv. Sci.'25]	57.13	66.78	71.42	60.48	41.63	61.41	62.87	53.99	57.91	62.49	71.95	**71.95**	66.74	73.91	79.23	81.30
DUSTED[AAAI'25]	46.53	59.12	64.00	61.63	51.10	66.38	67.08	63.68	43.08	62.76	64.35	58.60	60.31	69.29	75.76	75.94
STAIG[Nat. Com.'25]	50.36	64.09	63.02	57.19	59.45	69.51	71.60	67.42	52.88	67.02	65.46	60.33	45.17	65.56	59.25	58.48
SemST	**67.93**	**71.08**	**74.53**	**67.59**	**73.90**	**72.73**	**76.57**	**67.49**	**73.92**	**73.09**	**77.30**	68.29	**78.70**	**75.20**	**81.21**	**82.01**

Datasets	ME				MVC				HBC				MBA			
Methods	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1
STAGATE[Nat. Com.'22]	32.34	55.51	46.07	47.77	52.38	56.37	66.86	65.55	44.66	67.36	48.97	53.50	35.81	**72.49**	45.49	48.80
GraphST[Nat. Com.'23]	29.75	53.54	46.12	44.95	36.16	45.61	60.73	61.69	52.63	66.96	64.98	55.58	41.32	71.46	47.05	48.23
Spatial-MGCN(BIB'23)	44.54	59.91	56.79	57.38	53.32	59.89	74.31	74.40	65.68	**70.83**	64.67	65.71	48.34	68.03	44.12	44.87
GAAEST[Com. Biol.'24]	26.27	49.18	44.78	45.45	59.41	68.15	75.89	73.75	52.02	67.22	55.71	55.97	43.35	70.64	50.09	49.29
SEDR[Genome Med.'24]	27.70	51.32	46.32	46.89	52.71	63.16	67.77	66.17	43.16	67.30	50.34	54.19	40.36	71.37	46.86	50.19
MAFN[TKDE'24]	38.48	54.83	53.37	54.37	59.12	67.37	76.97	74.91	59.20	62.26	60.61	61.34	44.15	67.73	44.64	45.48
stDCL[Adv. Sci.'25]	34.32	59.34	51.68	50.69	49.54	63.49	66.94	62.17	55.73	67.80	58.19	56.21	42.05	70.75	48.24	45.06
DUSTED[AAAI'25]	26.23	52.15	40.03	39.67	58.06	66.30	71.25	71.33	47.81	65.78	48.92	52.40	35.86	71.13	43.75	47.31
STAIG[Nat. Com.'25]	27.20	51.02	46.68	47.62	58.26	69.58	71.91	70.61	57.86	69.43	59.82	63.19	33.35	70.61	44.12	46.68
SemST	**50.50**	**60.06**	**60.93**	**59.07**	**64.27**	**70.60**	**80.70**	**79.91**	**68.64**	68.23	**67.27**	**65.99**	**53.92**	70.12	**51.09**	**50.60**

Table 1: Quantitative comparison of clustering performance on eight spatial transcriptomics datasets using four evaluation metrics: ARI, NMI, ACC, and F1-score. **Bold** denotes the best result, and Underline denotes the second best.

Variants	151509				151510				151671				MBA			
	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1
w/o LLM	67.37	69.16	74.02	67.13	70.74	71.40	75.26	66.86	60.43	71.18	75.20	78.20	44.83	67.74	47.27	48.10
Random Emb.	59.95	59.45	69.61	60.74	63.16	60.84	73.23	54.40	50.56	62.79	67.99	71.35	41.72	65.35	42.97	41.65
Unrelated Emb.	65.24	68.23	71.60	61.85	65.96	68.50	74.67	66.39	60.11	72.62	72.67	76.71	43.89	68.20	46.79	48.63
BERT	63.27	65.71	68.34	60.19	69.25	70.30	75.50	67.28	60.05	70.67	73.93	77.40	42.96	68.35	45.86	46.61
Cross-Attention	62.40	56.75	64.68	66.26	66.07	67.48	72.36	64.70	63.14	74.03	75.64	78.87	47.69	67.45	48.50	49.25
Concat	69.25	68.91	72.99	62.50	67.43	69.88	73.69	65.79	73.21	67.64	80.23	80.12	46.64	68.05	47.01	47.97
Add	69.19	68.24	73.02	64.22	71.96	71.37	76.21	67.92	74.14	71.83	79.40	80.60	47.35	67.96	47.24	47.70
SemST	**73.90**	**72.73**	**76.57**	**67.49**	**73.92**	**73.09**	**77.30**	**68.29**	**78.70**	**75.20**	**81.21**	**82.01**	**53.92**	**70.12**	**51.09**	**50.60**

Table 2: Ablation study evaluating the contributions of semantic embeddings and the FSM module.

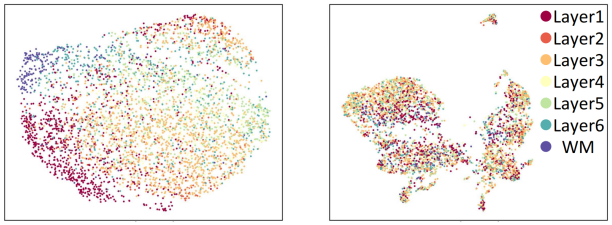
datasets. This confirms that the biologically informed semantic embeddings play a critical role in enabling the superior performance of SemST.

Quality of Semantic Embeddings. To further validate the importance of meaningful semantic information, we introduced three additional variants. "Random Emb." involved randomly initialized vectors with a mean of 0 and a variance of 0.85, a distribution chosen to approximate the LLM embedding. For "Unrelated Emb.", the LLM was prompted with entirely non-biological questions. The "BERT" variant merely employs a basic pre-trained BERT model (Devlin et al. 2019) as a plain text encoder, without incorporating any specialized domain knowledge. They almost consistently showed notably lower performance compared to the "w/o LLM" variant. This highlights the importance of using high-quality, biologically relevant semantic embeddings. Arbitrary or generic linguistic information may be ineffective or even detrimental. The UMAP visualization (Figure 3) demonstrates the superior quality of LLM-generated embeddings over that of BERT. The LLM captures a more smooth, biologically meaningful gradient across cortical layers (left), whereas the BERT embeddings are chaotic and lack clear structure (right). This illustrates that the LLM is capable of deriving deeper insights into gene symbols.

Effectiveness of FSM for Feature Fusion. We compared FSM module against several alternative fusion strategies for

Datasets	Methods	ARI	NMI	ACC	F1
		Before / After / Δ	Before / After / Δ	Before / After / Δ	Before / After / Δ
HBC	STAGATE [Nat. Com.'22]	44.66 / 58.16 / **+13.50**	67.36 / 68.80 / **+1.44**	48.97 / 58.77 / **+9.80**	53.50 / 61.17 / **+7.67**
	GraphST [Nat. Com.'23]	52.63 / 57.09 / **+4.46**	66.96 / 68.20 / **+1.24**	54.98 / 58.79 / **+3.81**	55.58 / 59.06 / **+3.48**
	Spatial-MGCN [BIB'23]	65.68 / 66.80 / **+1.12**	70.83 / 69.44 / −1.39	64.67 / 65.61 / **+0.94**	65.71 / 66.61 / **+0.90**
	GAAEST [Com. Biol.'24]	52.02 / 57.75 / **+5.73**	67.22 / 68.29 / **+1.07**	55.71 / 59.87 / **+4.16**	55.97 / 58.95 / **+2.98**
	SEDR [Genome Med.'24]	43.16 / 50.17 / **+7.01**	67.30 / 68.57 / **+1.27**	50.34 / 56.27 / **+5.93**	54.19 / 60.30 / **+6.11**
	MAFN [TKDE'24]	57.49 / 62.68 / **+5.19**	62.78 / 64.58 / **+1.80**	58.32 / 62.74 / **+4.42**	59.55 / 62.47 / **+2.92**
	stDCL [Adv. Sci.'25]	55.73 / 62.10 / **+6.37**	70.05 / 70.22 / **+0.17**	58.19 / 63.22 / **+5.03**	56.21 / 61.62 / **+5.41**
	DUSTED [AAAI'25]	47.81 / 50.95 / **+3.14**	65.78 / 66.20 / **+0.42**	48.92 / 52.24 / **+3.32**	52.40 / 56.24 / **+3.84**
	STAIG [Nat. Com.'25]	57.86 / 60.33 / **+2.47**	69.43 / 70.59 / **+1.16**	59.82 / 61.11 / **+1.29**	63.19 / 64.32 / **+1.13**
MBA	STAGATE [Nat. Com.'22]	35.81 / 38.79 / **+2.98**	72.49 / 72.11 / −0.38	45.49 / 46.64 / **+1.15**	48.80 / 50.05 / **+1.25**
	GraphST [Nat. Com.'23]	41.32 / 48.82 / **+7.50**	71.46 / 71.98 / **+0.52**	47.05 / 50.58 / **+3.53**	48.23 / 48.33 / **+0.10**
	Spatial-MGCN [BIB'23]	48.34 / 51.07 / **+2.73**	68.03 / 68.90 / **+0.87**	44.12 / 48.65 / **+4.53**	44.87 / 49.56 / **+4.69**
	GAAEST [Com. Biol.'24]	43.35 / 46.52 / **+3.17**	70.64 / 71.78 / **+1.14**	50.09 / 51.32 / **+1.23**	49.29 / 49.77 / **+0.48**
	SEDR [Genome Med.'24]	42.05 / 44.95 / **+2.90**	71.37 / 71.97 / **+0.60**	45.71 / 48.00 / **+2.29**	45.90 / 48.02 / **+2.12**
	MAFN [TKDE'24]	44.15 / 48.06 / **+3.91**	67.73 / 67.83 / **+0.10**	44.64 / 47.72 / **+3.08**	45.48 / 47.88 / **+2.40**
	stDCL [Adv. Sci.'25]	42.05 / 48.46 / **+6.41**	70.75 / 71.77 / **+1.02**	48.24 / 51.61 / **+3.37**	45.06 / 47.40 / **+2.34**
	DUSTED [AAAI'25]	35.86 / 38.30 / **+2.44**	71.13 / 71.68 / **+0.55**	43.75 / 45.68 / **+1.93**	47.13 / 48.38 / **+1.25**
	STAIG [Nat. Com.'25]	33.45 / 36.33 / **+2.98**	70.61 / 71.47 / **+0.86**	44.12 / 46.23 / **+2.11**	46.68 / 49.00 / **+2.32**

Table 3: Clustering performance comparison before and after integrating our proposed module into existing methods.



(a) Embeddings from the LLM (b) Embeddings from the BERT

Figure 3: UMAP visualization of semantic embeddings generated by the LLM and BERT on DLPFC slice #151508 dataset, colored according to manual annotations.

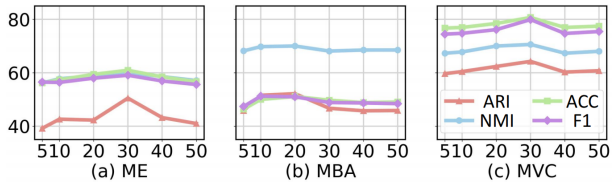


Figure 4: Impact of k_g (the number of top-expressed gene input to the LLM) on clustering performance across three datasets, evaluated by ARI, NMI, ACC, and F1.

integrating semantic embeddings with spatial features, including "Cross-Attention", "Concat" (concatenation), and "Add" (element-wise addition). SemST consistently outperforms all these alternatives, underscoring the superiority of FSM in enabling nuanced feature modulation.

Module Versatility. Based on the idea that biological semantics can universally enhance spatial representations, the FSM module is intended to function without reliance on any specific framework. To validate this plug-and-play versatility, we integrated semantic information via FSM into all baseline methods. As shown in Table 3, the results are compelling: the integration almost consistently yields significant performance improvements across nearly all methods and metrics, demonstrating the effectiveness of our approach in enriching spatial representations and improving performance on spatial domain identification.

Parameters Analysis

We examine how the number of top expressed genes (k_g) used as input affects clustering performance. As shown in Figure 4, SemST was evaluated on three datasets with k_g from 5 to 50. Performance generally improved with larger k_g, peaking at k_g=30 for ME and MVC, and k_g=20 for MBA, then declined. This suggests that a moderate-sized gene set best captures key biological signals, balancing information loss and gene redundancy.

Conclusion

In this work, we introduce SemST, a novel framework that addresses a critical limitation in spatial transcriptomics data clustering: the disregard for the biological semantics of gene symbols. By harnessing LLMs, SemST straightforwardly transforms the symbol sets of highly expressed genes into biologically rich semantic embeddings, which then function with the designed FSM module to modulate and guide spatial features. Remarkably, extensive experiments demonstrate that this simple yet effective approach leads to substantial performance improvements. By enabling genes to "speak" their biological roles within a spatial context, SemST offers a more profound and insightful perspective on the intricate tissue architecture.

Acknowledgements

The work was supported in part by the National Natural Science Foundation of China under grants 62476258 and 62522604, in part by the Natural Science Foundation of Hubei Province under grant 2025AFA113, in part by the Interdisciplinary Research Program of HUST under grant 2025JCY3024, and in part by the Fundamental Research Funds for National Universities, China University of Geosciences (Wuhan) under grant 2025XLB83.

References

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1): 25–29.

Biancalani, T.; Scalia, G.; Buffoni, L.; Avasthi, R.; Lu, Z.; Sanger, A.; Tokcan, N.; Vanderburg, C. R.; Segerstolpe, Å.; Zhang, M.; et al. 2021. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature methods*, 18(11): 1352–1362.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, W.; Zhang, P.; Tran, T. N.; Xiao, Y.; Li, S.; Shah, V. V.; Cheng, H.; Brannan, K. W.; Youker, K.; Lai, L.; et al. 2025. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 1–15.

Chen, Y.; and Zou, J. 2025. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature biomedical engineering*, 9(4): 483–493.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dong, K.; and Zhang, S. 2022. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1): 1739.

Du, J.; Yang, Y.-C.; An, Z.-J.; Zhang, M.-H.; Fu, X.-H.; Huang, Z.-F.; Yuan, Y.; and Hou, J. 2023. Advances in spatial transcriptomics and related data analysis strategies. *Journal of translational medicine*, 21(1): 330.

He, X.; Tang, C.; Liu, X.; Li, C.; An, S.; and Li, Z. 2024. Heterogeneous graph guided contrastive learning for spatially resolved transcriptomics data. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8287–8295.

Hou, W.; and Ji, Z. 2024. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature methods*, 21(8): 1462–1465.

Hu, J.; Li, X.; Coleman, K.; Schroeder, A.; Ma, N.; Irwin, D. J.; Lee, E. B.; Shinohara, R. T.; and Li, M. 2021. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11): 1342–1351.

Hu, M.; Alkhairy, S.; Lee, I.; Pillich, R. T.; Fong, D.; Smith, K.; Bachelder, R.; Ideker, T.; and Pratt, D. 2025. Evaluation of large language models for discovery of gene set function. *Nature methods*, 22(1): 82–91.

Huang, C. H. 2024. QuST-LLM: Integrating Large Language Models for Comprehensive Spatial Transcriptomics Analysis. arXiv:2406.14307.

Lampert, M.; Blöcker, C.; Scholtes, I.; and Grün, D. 2024. Cell-Type Prediction in Spatial Transcriptomics Data using Graph Neural Networks. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*.

Li, J.; Chen, S.; Pan, X.; Yuan, Y.; and Shen, H.-B. 2022. Cell clustering for spatial transcriptomics data with graph neural networks. *Nature Computational Science*, 2(6): 399–408.

Liu, J.; Yang, M.; Yu, Y.; Xu, H.; Li, K.; and Zhou, X. 2024a. Large language models in bioinformatics: applications and perspectives. arXiv:2401.04155.

Liu, T.; Fang, Z.-Y.; Zhang, Z.; Yu, Y.; Li, M.; and Yin, M.-Z. 2024b. A comprehensive overview of graph neural network-based approaches to clustering for spatial transcriptomics. *Computational and Structural Biotechnology Journal*, 23: 106–128.

Long, Y.; Ang, K. S.; Li, M.; Chong, K. L. K.; Sethi, R.; Zhong, C.; Xu, H.; Ong, Z.; Sachaphibulkij, K.; Chen, A.; et al. 2023. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1): 1155.

Longo, S. K.; Guo, M. G.; Ji, A. L.; and Khavari, P. A. 2021. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, 22(10): 627–644.

Marx, V. 2021. Method of the Year: spatially resolved transcriptomics. *Nature methods*, 18(1): 9–14.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rodriques, S. G.; Stickels, R. R.; Goeva, A.; Martin, C. A.; Murray, E.; Vanderburg, C. R.; Welch, J.; Chen, L. M.; Chen, F.; and Macosko, E. Z. 2019. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434): 1463–1467.

Satija, R.; Farrell, J. A.; Gennert, D.; Schier, A. F.; and Regev, A. 2015. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5): 495–502.

Ståhl, P. L.; Salmén, F.; Vickovic, S.; Lundmark, A.; Navarro, J. F.; Magnusson, J.; Giacomello, S.; Asp, M.; Westholm, J. O.; Huss, M.; et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82.

Wang, B.; Luo, J.; Liu, Y.; Shi, W.; Xiong, Z.; Shen, C.; and Long, Y. 2023. Spatial-MGCN: a novel multi-view graph convolutional network for identifying spatial domains with attention mechanism. *Briefings in Bioinformatics*, 24(5): bbad262.

Wang, T.; Zhu, H.; Zhou, Y.; Ding, W.; Ding, W.; Han, L.; and Zhang, X. 2024. Graph attention automatic encoder based on contrastive learning for domain recognition of spatial transcriptomics. *Communications Biology*, 7(1): 1351.

Xu, H.; Fu, H.; Long, Y.; Ang, K. S.; Sethi, R.; Chong, K.; Li, M.; Uddamvathanak, R.; Lee, H. K.; Ling, J.; et al. 2024. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine*, 16(1): 12.

Yang, Y.; Cui, Y.; Zeng, X.; Zhang, Y.; Loza, M.; Park, S.-J.; and Nakai, K. 2025. STAIG: Spatial transcriptomics analysis via image-aided graph contrastive learning for domain exploration and alignment-free integration. *Nature Communications*, 16(1): 1067.

Yang, Y.; Hossain, M. Z.; Li, X.; Rahman, S.; and Stone, E. 2024. Spatial Transcriptomics Analysis of Zero-Shot Gene Expression Prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 492–502. Springer.

Yu, Z.; Lu, Y.; Wang, Y.; Tang, F.; Wong, K.-C.; and Li, X. 2022. Zinb-based graph embedding autoencoder for single-cell rna-seq interpretations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 4671–4679.

Yu, Z.; Yang, Y.; Chen, X.; Wong, K.-C.; Zhang, Z.; Zhao, Y.; and Li, X. 2025. Accurate Spatial Heterogeneity Dissection and Gene Regulation Interpretation for Spatial Transcriptomics using Dual Graph Contrastive Learning. *Advanced Science*, 12(3): 2410081.

Zahedi, R.; Ghamsari, R.; Argha, A.; Macphillamy, C.; Beheshti, A.; Alizadehsani, R.; Lovell, N. H.; Lotfollahi, M.; and Alinejad-Rokny, H. 2024. Deep learning in spatially resolved transcriptomics: a comprehensive technical view. *Briefings in Bioinformatics*, 25(2): bbae082.

Zeng, Z.; and Du, H. 2023. Revolutionizing single cell analysis: The power of large language models for cell type annotation. arXiv:2304.02697.

Zhao, E.; Stone, M. R.; Ren, X.; Guenthoer, J.; Smythe, K. S.; Pulliam, T.; Williams, S. R.; Uytingco, C. R.; Taylor, S. E.; Nghiem, P.; et al. 2021. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature biotechnology*, 39(11): 1375–1384.

Zhou, X. 2022. Graphing cell relations in spatial transcriptomics. *Nature Computational Science*, 2(6): 354–355.

Zhu, J.; Li, Y.; Tang, Z.; and Chang, C. 2025. DUSTED: Dual-Attention Enhanced Spatial Transcriptomics Denoiser. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1219–1227.

Zhu, Y.; He, X.; Tang, C.; Liu, X.; Liu, Y.; and He, K. 2024. Multi-view Adaptive Fusion Network for Spatially Resolved Transcriptomics Data Clustering. *IEEE Transactions on Knowledge and Data Engineering*.